



Medical Text Mining: A DLI-2 Status Report

Acknowledgement:
NSF DLI2, NIH/NLM

Hsinchun Chen
McClelland Professor,
Director,
Artificial Intelligence Lab and
Hoffman eCommerce Lab
The University of Arizona

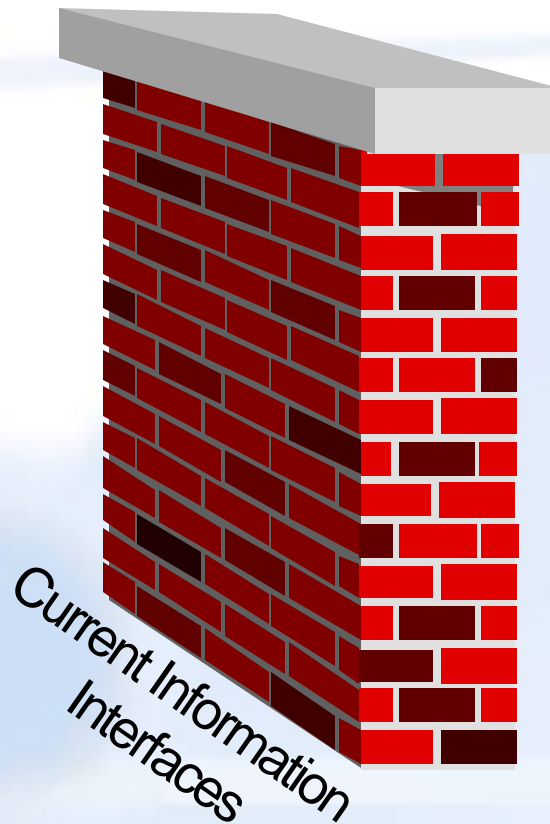
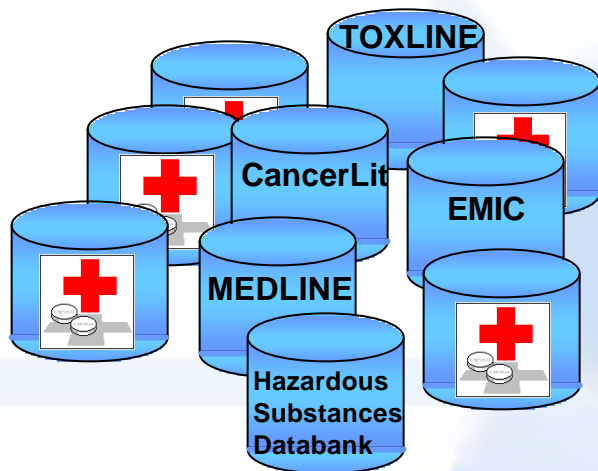


Artificial Intelligence Lab

The University of Arizona The University of Arizona The University of Arizona The University of Arizona The University of Arizona

The Medical Information Gap

Heterogeneous
Medical
Literature Databases
and the Internet



Medical
Professionals
& Users





Research Questions in Medical Text Mining

- How can linguistic parsing and statistical analysis techniques help extract medical terminology and the relationships between terms?
- How can medical and general ontologies help improve extraction of medical terminology?
- How can linguistic parsing, statistical analysis, and ontologies be incorporated in customizable retrieval interfaces?

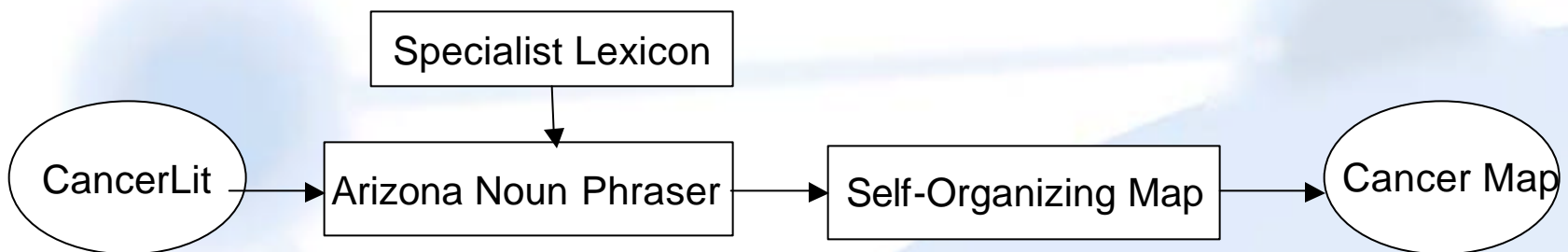


Medical Text Mining Research Update

- Cancer Map: Knowledge Map for Cancer Researchers
- Cancer Spider: Client-based Meta Cancer Search Agent
- Chinese MED Map: Multilingual Research in Medical Informatics



Cancer Map System Architecture



- Linguistic analysis: Arizona Noun Phraser and UMLS Specialist Lexicon
- Kohonen Self-Organizing Map (SOM): Topic and document categorization
- Multi-layered graphical display of important cancer concepts supports *browsing* of cancer literature (1M+ CancerLit documents)
- Presents 21,000 cancer topics on 1180 maps organized in 5 layers



Artificial Intelligence Lab

The University of Arizona The University of Arizona The University of Arizona The University of Arizona The University of Arizona

Browsing Cancer Map

1 Visual Site Browser

2 Top level map

3 Diagnosis, Differential

4 Brain Neoplasms

5 Brain Tumors

The final window (5) displays the 'OOHAY Concept & Document Server' interface. It shows a 'List of Found Documents' with 'Total Found Documents: 83'. The selected document is 'DOC UI: 99317761'.

Authors: Bloem BR | de Roos MA | de Beaufort AJ | Brouwer OF

Title: The stumbling toddler

Source: Ned Tijdschr Geneesk; 143(23):1185-8, 1999

Central Concepts: Brain Neoplasms | Cerebellar Ataxia | Gait | Medulloblastoma

MeSH: Anti-Anxiety Agents, Benzodiazepine | Case Report | Child, Preschool | Clonazepam Dipotassium | Diagnosis, Differential | Diarrhea | English Abstract | Female | Human | Male | Neurologic Examination | Recovery of Function | Tomography, X-Ray Computed | Treatment Outcome | Vertigo | Virus Diseases

Abstract: Four previously healthy children, two boys aged 6 and one boy and one girl aged 4 more or less acutely developed a stumbling gait. The causes varied from benign such as postviral acute cerebellar ataxia and benign paroxysmal vertigo to potentially life-threatening such as intoxication with benzodiazepines and medulloblastoma. Treatment led to complete or partial recovery. (Sub)acute balance disorders in previously healthy children can be due to cerebellar ataxia, vestibular disorders, and abnormal proprioception. Ancillary investigations are warranted in case of gradually developing ataxia, accompanying neurological deficits, suspicion of intoxication, recurrent or familial ataxia, no spontaneous remission or even progression. In children with an isolated cerebellar ataxia without these features, ancillary investigations may be avoided, although in such cases careful follow-up remains necessary.

DOC UI: 96272136



Cancer Map User Study

- Compare topic hierarchies on Cancer Map vs. topic hierarchies on MeSH cancer subtree
- 30 cancer researchers from Arizona Cancer Center as subjects
- Future work in interface comparison: size, proximity, and layers (Topic Island)



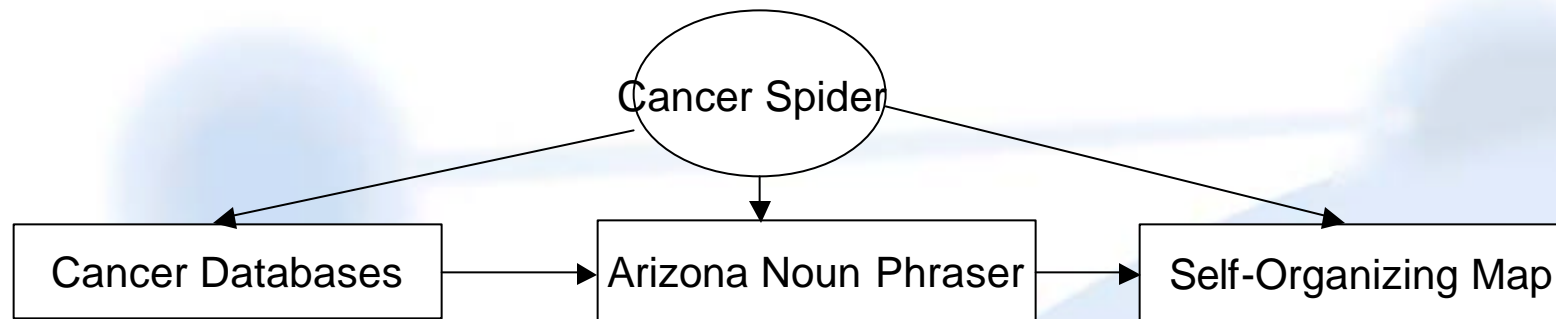
Cancer Map User Study Result

C: Cancer Map M: MeSH	First level	Second level (overlap)	Second level (non-overlap)	Third level
Recall Comparison	C: 0.557	C: 0.765	C: 0.859	C: 0.839
	M: 0.466	M: 0.113	M: 0.466	M: 0.459
	P = 0.049	P = 0.00	P = 0.000	P = 0.003
Precision Comparison	C: 0.926	C: 0.826	C: 0.829	C: 0.863
	M: 0.956	M: 0.608	M: 0.904	M: 0.917
	P = 0.591	P = 0.104	P = 0.459	P = 0.808

- Cancer Map was comparable to MeSH cancer subtree in *perceived topic precision* at each level and was significantly better than MeSH in *perceived topic recall* at all levels.
- Cancer Map and MeSH are complementary in topic suggestion.

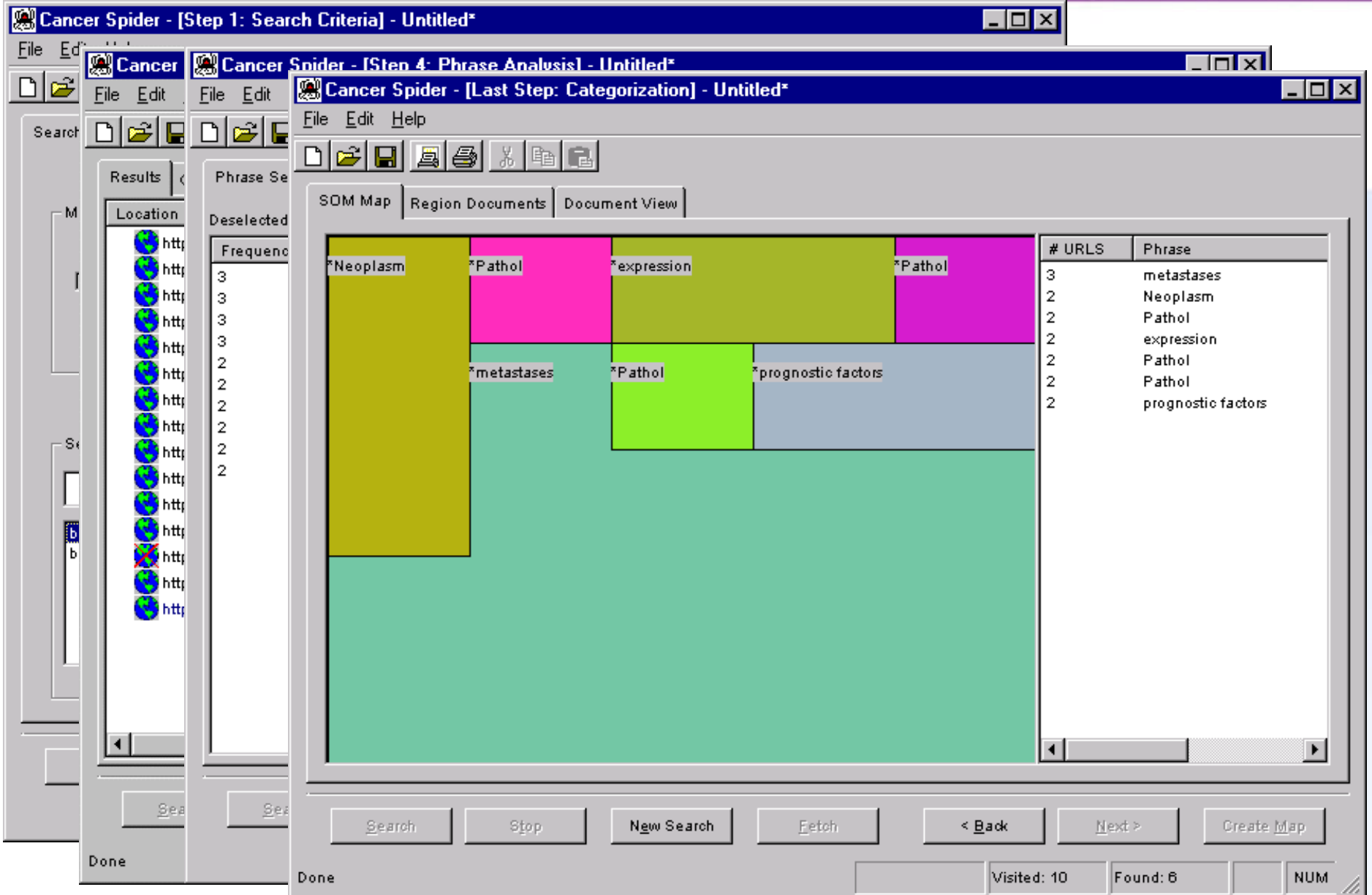


Cancer Spider System Architecture



- Cancer databases: CancerLit, PDQ, Medline
 - Linguistic analysis: Arizona Noun Phraser and UMLS Specialist Lexicon
 - Kohonen Self-Organizing Map (SOM): Topic and document categorization
-
- Meta searching: Search multiple cancer databases
 - Iterative, dynamic, personalized medical topic/theme summarization

The University of Arizona The University of Arizona The University of Arizona The University of Arizona The University of Arizona





Cancer Spider User Study

- Compare Cancer Spider vs. NLM Gateway (NLM's one-stop MED portal with MeSH thesaurus)
<http://gateway.nlm.nih.gov/gw/Cmd?GMBasicSearch>
- 30 cancer researchers from Arizona Cancer Center as subjects; 2 expert evaluators
- Performance measures: document recall, precision, F, time spent, # of documents



Cancer Spider User Study Result

	Sample size	CancerSpider		NLM Gateway		P-Value
		Mean	Variance	Mean	Variance	
Precision	30	0.803	0.117	0.826	0.122	0.7572
Recall	30	0.533	0.112	0.539	0.121	0.9523
F-measure	30	0.612	0.105	0.622	0.110	0.9056

- Cancer Spider and NLM Gateway users achieved similar performances in precision, recall, and F measure.



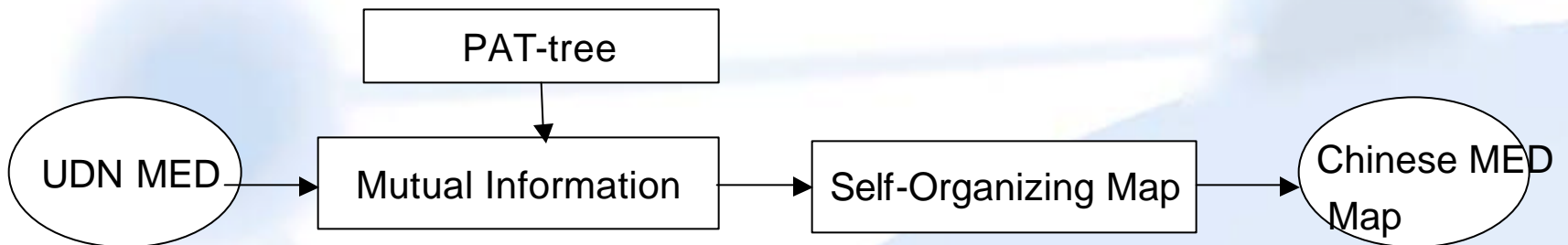
Cancer Spider User Study Result

	Sample size	CancerSpider		NLM Gateway		P-Value
		Mean	Variance	Mean	Variance	
Time (in minutes)	30	10.22	15.64	14.00	23.18	0.0003
No. of documents browsed	30	4.533	3.586	6.233	12.461	0.0067

- Cancer Spider users spent less time and browsed fewer documents.
- The two tools are complementary in functionalities and documents suggested.



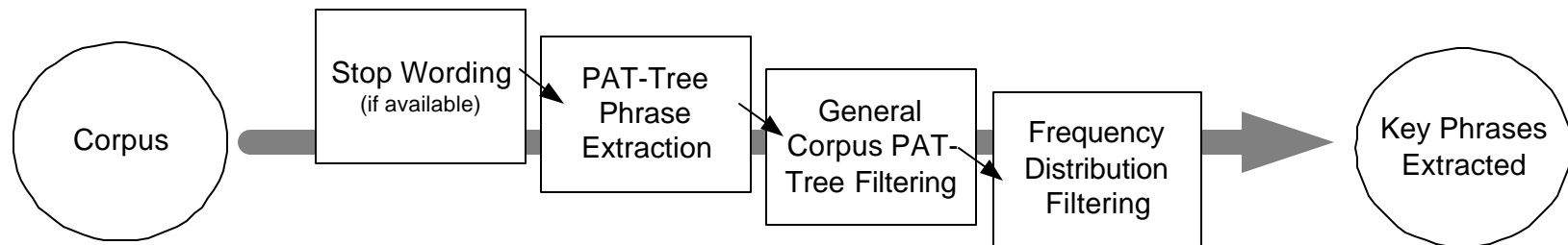
Chinese MED Map System Architecture



- Statistic-based indexing: Mutual Information and PAT-tree
- Kohonen Self-Organizing Map (SOM): Topic and document categorization
- Multi-layered graphical display of important health-related topics supports *browsing* of health-related news articles (70K+ United Daily News, UDN MED)
- User study underway



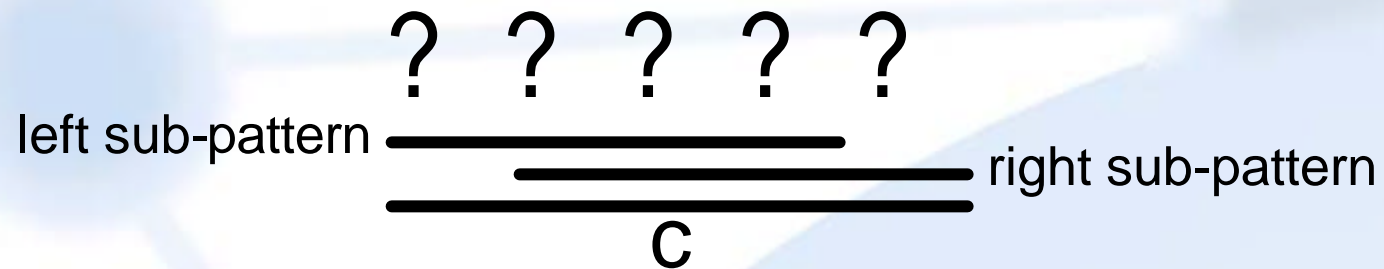
Chinese Medical Indexing Architecture



- Stop wording breaks long sentence into smaller chunks to reduce noise
- “Updateable” mutual information technique improves precision of extraction
- General PAT-tree filter general phrases
- Frequency distribution filtering distills less useful terms



Mutual Information (MI) estimator



$$MI_c = \frac{f_c}{f_{left} + f_{right} + f_c}$$

Note: $|A \neq B| \neq |A| \neq |B| \neq |A \neq B|$

$$MI_c = \log_2 \frac{\Pr(c)}{\Pr(left) \Pr(right)}$$

Independent event: $\Pr(A \neq B) \neq \Pr(A) \Pr(B)$

$MI_c = 0$ no correlation

$MI_c = 1$ perfect correlation



Artificial Intelligence Lab

The University of Arizona The University of Arizona The University of Arizona The University of Arizona The University of Arizona

List of extracted phrases

Sample Medical Abstract

The screenshot shows two windows from a Windows 95 desktop. The left window, titled 'C:\temp\list.html - Micros...', displays a list of extracted phrases with associated numerical values. The right window, titled 'M:\data\stic\report_med.html - Microsoft Internet Explorer', displays a sample medical abstract in Chinese. The abstract text is highlighted with various colors (green, blue, red) to indicate different types of extracted terms.

List of extracted phrases (Left Window):

Phrase	Value 1	Value 2	Value 3	Value 4	Value 5
荷爾蒙接受體	1.000	3	3	3	
蛇毒抑制成分	1.000	5	5	5	
陰部潰瘍患者	1.000	3	3	3	
最大運動量下	0.833	10	12	10	
單球性白血病	0.600	3	3	5	
壺腹周圍腫瘤	1.000	3	3	3	
復健工作訓練	1.000	4	4	4	
游離甲狀腺素	1.000	6	6	6	
發生急性轉化	1.000	3	3	3	
短小包膜條蟲	1.000	3	3	3	
結核菌素試驗	1.000	3	3	3	
腎源前列腺素	1.000	3	3	3	
視覺誘發反應	1.000	5	5	5	
進入胃癌細胞	1.000	4	4	4	
鈣促化劑刺激	1.000	4	4	4	
間接連續刺激	1.000	3	3	3	
嗜中性白血球	0.417	5	5	12	
嗜中性球吞噬	1.000	3	3	3	

Sample Medical Abstract (Right Window):

冠狀動脈心臟病之診斷，傳統上常借助於運動心電圖試驗，然其診斷靈敏度仍偏低，本研究之目的，乃探討以運動式核子心臟造影術診斷冠狀動脈心臟病之可行性，以及乙型阻斷劑對本檢查之影響，檢查對象包括18名（男性7名，女性11名），正常者及50名（男43名，女7名）冠狀動脈心臟病患者，受檢者取坐姿，在醫用腳踏車上運動，運動之前及最大運動量時各作一次第一循環核子心臟造影術檢查，所用攝影機為電腦化多晶體閃爍型，運動前、運動中及運動後10分鐘內均監視並記錄心電圖（v5導程）變化，血壓則每2分鐘測量一次，核子心臟檢查所得心室功能等數據與冠狀動脈造影之病變相互比較，結果顯示：經臨床種種檢查包括心導管及冠狀動脈攝影證實為正常者（18名），其靜態及最大運動量下之左心室射出分率（left ventricular ejection fraction, lvef）平均值及標準偏差分別為：63.9%及70.10%（p<0.01），50名冠狀動脈病患者，其靜態及最大運動量下之lvef分別為：1.條冠狀動脈阻塞（14名）：55.19%及59.18%（p>0.05）；2.條血管阻塞（18名）：53.16%及56.18%（p>0.05）；3.條血管阻塞（18名）：53.11%及57.13%（p>0.05），核子心臟造影於冠狀動脈病之診斷依據為：最大運動量下lvef之上昇小於5%或最大運動量下出現新的局部心壁運動異常，依據此，核子心臟造影術之診斷靈敏度為：1.條血管阻塞 10 / 14 (71.4%) ; 2.條血管阻塞 15 / 18 (83.3%) ; 3.條血管阻塞 15 / 18 (83.3%)

Extracted medical terms appear in a medical abstract

Topic Map

- 呼吸道
- 憂鬱
- 糖尿病
- 脊髓
- 腫瘤
- 荷爾蒙
- 關節
- 子宮內膜
- 子宮頸
- 安眠藥
- 小兒麻痺
- 心臟病
- 愛滋病
- 抗生素
- 整形外科
- 染色體
- 潰瘍
- 癡呆
- 癲癇
- 白血球
- 精神病患
- 耳鼻喉
- 肝炎
- 脊椎
- 膀胱

UdnMed

Topic Map

直腸癌

- UdnMed
- 呼吸道
- 憂鬱
- 糖尿病
- 脊髓
-

紅血球

荷爾蒙

直腸癌

子宮內膜

臨下垂體

Authors: ◎梁金銅台大醫院大腸直腸外科主治醫師

Title: 大腸直腸癌

Source: 《聯合報》

Abstract: 近年來因經濟的起飛，人口結構的老化，加上生活型態的改變，西式飲食的盛行，導致台灣地區之大腸直腸癌發生率及死亡率節節上揚。就死亡率而言，大腸直腸癌目前已是台灣地區因惡性腫瘤死亡人口的第三位，均僅次肝癌及肺癌。不論在我國或先進國家，大腸直腸癌已是今日公共衛生重要的一環。近年來有關大腸直腸癌的流行病學研究甚多，其中較具體的結論是遺傳與飲食。我們大概可以說家族一等親中若有人得到大腸直腸癌，則其一生中得相同癌症的機會約為一般人的三倍。目前公認纖維質食物攝取太少，以及攝取太多的肉類，由於會導致大便通過大腸的平均時拉長，所以致癌的機會也會大增。就大腸直腸癌病變而言，



Future Research Directions

- Text mining + MED Ontologies
- Information visualization for medical informatics
- Multilingual medical informatics: text mining + Multilingual UMLS
- GeneScene: Gene pathway analysis and visualization



Research Announcement:

- ICADL2001, International Conference of Asian DL, Bangalore, India, 12/10-12/12/ 2001 (paper due: 7/15/2001)
- CFP, Special Topics Issue of DSS: “Web Retrieval and Mining”, 7/13/2001
- CFP, Special Topics Issue of DSS: “Digital Government: Technologies and Practices”, 8/14/2001
- CFP, Special Topics Issue of JASIST: “Web Retrieval and Mining”, 9/14/2001



Artificial Intelligence Lab

The University of Arizona The University of Arizona The University of Arizona The University of Arizona The University of Arizona

For Project Information:

<http://ai.bpa.arizona.edu>

Hchen@bpa.arizona.edu

For Medical Demos:

<http://www.HelpfulMED.com>